Data Integration with Cloud Data Fusion

Course 4342 – 16 Hours

Overview

This 2-day course introduces learners to Google Cloud's data integration capability using Cloud Data Fusion. In this course, we discuss challenges with data integration and the need for a data integration platform (middleware). We then discuss how Cloud Data Fusion can help to effectively integrate data from a variety of sources and formats and generate insights. We take a look at Cloud Data Fusion's main components and how they work, how to process batch data and real time streaming data with visual pipeline design, rich tracking of metadata and data lineage, and how to deploy data pipelines on various execution engines.

On Completion, Delegates will be able to

- Identify the need of data integration,
- Understand the capabilities Cloud Data Fusion provides as a data integration platform,
- Identify use cases for possible implementation with Cloud Data Fusion,
- List the core components of Cloud Data Fusion,
- Design and execute batch and real time data processing pipelines,
- Work with Wrangler to build data transformations
- Use connectors to integrate data from various sources and formats,
- Configure execution environment; Monitor and Troubleshoot pipeline execution,
- Understand the relationship between metadata and data lineage

Who Should Attend

- Data Engineer
- Data Analysts

Prerequisites

To get the most out of this course, participants are encouraged to have:

- Completed "Big Data and Machine Learning Fundamentals"

Course Contents

- Introduction
  - Course Introduction
- Introduction to data integration and Cloud Data Fusion
  - Data integration: what, why, challenges
  - Data integration tools used in industry
  - User personas
  - Introduction to Cloud Data Fusion
  - Data integration critical capabilities
  - Cloud Data Fusion UI components
- Building pipelines
  - Cloud Data Fusion architecture
  - Core concepts
  - Data pipelines and directed acyclic graphs (DAG)
  - Pipeline Lifecycle
  - Designing pipelines in Pipeline Studio
- Designing complex pipelines
  - Branching, Merging and Joining
  - Actions and Notifications
  - Error handling and Macros
  - Pipeline Configurations, Scheduling, Import and Export
- Pipeline execution environment
  - Schedules and triggers
  - Execution environment: Compute profile and provisioners
  - Monitoring pipelines
- Building Transformations and Preparing Data with Wrangler
  - Wrangler
  - Directives
  - User-defined directives
- Connectors and streaming pipelines
  - Understand the data integration architecture.
  - List various connectors.
  - Use the Cloud Data Loss Prevention (DLP) API.
  - Understand the reference architecture of streaming pipelines.
  - Build and execute a streaming pipeline.
- Metadata and data lineage
- Summary